Львівський національний університет імені Івана Франка

На правах рукопису

Волощук Орест Романович

УДК 004.853+004.855.5

Методи кластеризації на великих масивах даних

01.05.03 — математичне та програмне забезпечення обчислювальних машин і систем

Магістерська робота

Науковий керівник

Годич Олесь Васильович,

кандидат фізико-математичних наук, доцент

Львів — 2011

ЗМІСТ

ВСТУП

Актуальність теми. Сьогодні все частіше виникають задачі, так чи інакше пов'язані із розбиттям масиву об'єктів на групи за певними критеріями. Із розвитком обчислювальної техніки збільшуються також об'єми баз даних, що можуть піддаватись такому аналізу. Виникає необхідність створення оптимальних алгоритмів обробки великих масивів даних.

Мета і завдання дослідження. Метою дослідження є розвиток методики кластеризації великих масивів даних. Для досягнення цієї мети були сформульовані та вирішені такі основні завдання:

— провести детальний аналіз та дослідити ефективність різноманітних алгоритмів кластеризації;

— здійснити контроль якості результатів роботи алгоритмів;

— розробити методи та виявити можливі оптимізації, що дозволять прискорити виконання задачі кластеризації

Об'єкт дослідження. Об'єктом дослідження є методи ієрархічні та плоскі алгоритми кластеризації даних.

РОЗДІЛ 1
ОГЛЯД СТАНУ ПРОБЛЕМИ ТА ОСНОВНІ ПОНЯТТЯ

Термін "кластеризація"вперше використав Тріон у [1]

Кластеризацією називається процес розбиття певної заданої множини об'єктів на підмножини, що називаються *кластерами*, так, щоб в одній підмножині (кластері) опинились об'єкти, якимось чином схожі між собою. Ці об'єкти також називають *спостереженнями*. Кожне спостереження характеризується певним скінченним набором атрибутів. Якщо у кожного об'єкта є точно n атрибутів, і якщо всі ці атрибути у якийсь спосіб приведено до числового вигляду, то можна розцінювати такий масив даних як набір точок у n-вимірному просторі.

1.0.1. Задачі кластеризації. За останні кілька десятиліть виникла ціла низка задач, розв'язання яких на певному етапі вимагає кластеризації вхідних даних. Ці задачі походять із різноманітних сфер людської діяльності – соціології, медицини, математики, і навіть із комерційної діяльності. Кластеризація зазвичай застосовується при розпізнаванні образів, виявленні прихованих закономірностей даних, та ін.

Для просторів об'єктів, розмірність яких не перевищує 2, людина здатна провести якісну кластеризацію вручну. Для цього достатньо представити вхідні дані як масив точок на двовимірній площині або одновимірній прямій (в залежності від розмірності простору). Таким чином зображені дані піддаються швидкому візуальному аналізу, що дозволяє ефективно встановити основні кластери (або ж неможливість кластеризації даного набору). Для просторів більшої розмірності така методика кластеризації недоступна. Це пов'язано із неможливістю наглядно зобразити вхідні дані у ви-

гляді геометричних об'єктів.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Tryon, R. C. (1939) *Cluster Analysis*. Ann Arbor: Edwards Brothers.